一种面向中文本体模式的本体对齐框架*

王 汀 高 迎 刘经纬
(首都经济贸易大学信息学院 北京 100070)

摘要: 【目的】现有的本体对齐方法往往忽视中文概念的语序敏感和一词多义的语义特征。本文提出一种基于同义词词林和序列比对算法的大规模中文本体映射模型。【方法】采用基于改进的同义词词林相似度算法计算简单词元的语义相似度。并利用基于改进同义词词林与序列比对相融合的算法度量未登录词之间的语义相似度。【结果】在由 DBpedia(中文版)、百度百科和互动百科知识库所构建的测试语料上的关联映射实验结果表明, 该模型的准确率、召回率和综合评价指标平均分别达到约 97.5%、87.8%和 92.1%。【局限】本模型仅专注于对中文本体概念的元素级相似度度量, 并未考虑本体属性和实例对于概念等价关系的影响因素。【结论】在面向中文网络百科的大规模开放语义数据集上的评测结果证明, 该模型的总体性能明显优于现有算法。

关键词: 中文关联数据 同义词词林 序列比对 本体映射 相似度计算
分类号: G353.1

1 引 言

语义 Web 的愿景是建立“数据之网”(Web of Data), 以使机器能够理解网络上的语义信息[1]。本体作为语义 Web 的核心元素, 是描述特定领域共享概念的形式化、规范化说明[2], 是实现网络知识共享和语义互操作的基础。目前关联数据(Linked Open Data, LOD)[3]的研究工作主要集中在面向实例级别(Level of Instances)上展开[4-5], 同时, 由于不同本体之间存在异构性, 导致本体间的重用和共享变得困难。因此, 作为关联数据的基础和前提, 面向模式级别(Schema-Level)的关联数据构建研究亦很重要[6]。

本体映射(Ontology Mapping)作为模式级的关联数据构建典型场景已被广泛研究, 其任务就是要发现异构本体或数据源(LOD Datasets)之间的概念语义关联。而随着语义网的蓬勃发展, 中文描述的大规模本体和知识库也越来越多地被构建和共享出来。同时, 由于文化和背景的原因, 目前大规模中文关联数据网络的构建研究尚处于初级阶段, 更缺乏成熟的面向模式级别的大规模中文关联数据模型。因此, 为了解决在关联数据网络中的中文本体语义互操作和共享问题, 本文面向本体模式层面, 提出一种新的大规模中文本体映射模型。

2 相关工作

国内外研究人员已提出多种映射方法和典型系统。Melnik 等[7]提出一种结构级本体映射算法: Similarity Flooding, 利用本体的概念体系构造相似度传播图, 并对概念之间的相似度进行传播和修正。Cohen 等[8]分析基于编辑距离和基于 Token 的几种典型元素级相似度计算算法, 并对几种算法的性能进行评测。Giunchiglia 等[9]提出基于语言学方法, 并引入共享知识词典(如: WordNet[10]), 利用语言关系进行语义关系发现。Isaac 等[11]提出一种实例级本体映射算法, 根据本体概念的公共实例数量来度量概念的相似度。Nikolov 等[12]基于工作流技术提出链接数据的框架 KnoFuss, 利用本体库中概念之间的层次关系选择最合适的匹配方法以及匹配参数。Zhong 等[13]提出

通讯作者: 王汀, ORCID: 0000-0003-2481-2890, E-mail: wangting@cueb.edu.cn。

*本文系首都经济贸易大学科研项目“基于数据场和序列比对的中文关联数据构建研究”(项目编号: 00791554410264)和北京市哲学社会科学项目“‘互联网+’环境下北京公共信息流动机制及协同获取模式研究”(项目编号: 16srb021)的研究成果之一。

RiMOM 系统，该系统基于本体实例、概念名称以及本体结构等特征的多策略映射方式，并通过引入普适的场论思想，使其适用于大规模本体的映射任务。Jain 等[6]发布了 BLOOMS 系统，该系统基于 Bootstrapping 方法并采用 Wikipedia 顶层分类树作为相似度计算知识库，从而进行 LOD 环境中的面向本体模式的链接构建。但是上述系统均只能针对和处理英文描述的语义数据集的本体模式映射任务。

近年来，越来越多的学者开始关注中文本体及其关联数据的构建工作。特别是在面向本体模式级别(即：本体映射)的中文关联数据网络建设层面上，李佳等[14]提出一种基于知网(HowNet)[15]的元素层概念相似度计算的方法并实现中文本体映射系统，但该系统忽视了中文普遍存在的"语序敏感"和"一词多义"现象[16]，因此在面对大规模本体映射任务时，其在关联数据环境中的适用性有待验证。基于《同义词词林》(扩展版)[17]，田久乐等[18]提出一种中文词语语义相似度计算算法，但并未涉及对于中文未登录词的相似度计算处理方式，其成果也未在实际的大规模关联数据网环境下应用。

除此之外，也有很多面向实例级别的典型关联数据系统。Silk[19-20]是一个在不同数据集之间实现链接的框架，其设计了一种声明式语言，用户可以对两个数据集之间的链接进行配置，包括链接的类型和链接的条件，并且可以实现远程数据集与本地数据集的链接。Hassanzadeh 等[21]提供了一个通用和可扩展的框架 LinQL，其中集成了很多已有的发现关联的方法。该框架的目的是帮助用户选择最适合的数据集的关联方法。同时，还支持基于关系数据库进行发布的 RDF 数据，例如使用 D2RQ 或 Virtuoso 发布的关系数据。Wang 等[5]提出基于中文百科的分类体系 DMOZs，抽取概念之间的层次关系并获取含有 Infobox 的词条 Web 页面中的概念属性及百科词条实例，最终建立起基于百度百科和互动百科的两大中文大规模本体库，并根据简单的关键字匹配策略，与 DBpedia 建立起实例间的共指关系。Niu 等[4]将百度百科[22]、互动百科[23]以及中文维基百科[24-25]进行语义集成，并开发出基于中文描述的实例级关联数据应用系统 Zhishi.me。为了实现在关联数据网络环境中的知识共享、重用和语义互操作，跨语言的本体链接和映射就成为必须要解决的问题。Wang 等[26]提出采用概念标注方法，借助少量的跨语言链接和内部链接种子来丰富内部链接，并在此基础上采用回归学习模型来预测中英文维基百科之间潜在的跨语言链接。但是上述系统均只涉及实例之间的关联关系构建,而缺乏对于本体模式层面上的链接获取和发现。

综上所述，目前发布在 Web 上的中文大规模本体仍然较少，且存在较大的异构性，而现有的中文本体映射系统在面对大规模本体映射任务时，效率较低且可用性不高。同时，仍缺乏针对中文语言描述且适应 LOD 环境的大规模本体映射系统。因此，本文基于《同义词词林》(扩展版)和序列比对思想，提出一种新的中文本体映射模型。该模型可以有效解决中文概念相似度计算时出现的语序敏感和一词多义问题。在基于中文网络百科构建的大规模本体测试集上的实验结果表明，该系统可以获得高于前人工作的总体性能。

3　问题定义

《同义词词林》(TongYiCiCiLin，TYCCL)(扩展版)中已收录的词汇称为简单词元。在中文本体映射系统中，简单词元与未登录词都对应于本体概念。本文将简单词元称为原子概念(Atom Concept，AC)，将未登录词统称为组合概念(Component Concept，CC)，并约定组合概念由若干个原子概念的线性排列组合而成。下面给出问题的定义：

定义 1：本体映射：两个待映射本体 O_s、O_t，对于 O_s 中的概念 C_s，在 O_t 中找到与其语义相同或接近的概念 C_t，有映射函数 $map:O_s\rightarrow O_t$：

对于 $\forall C_s\in O_s$，$\forall C_t\in O_t$，若 $sim(C_s, C_t)>t$；则有 $map(C_s)=C_t$

$sim(C_s, C_t)$为 C_s 和 C_t 的相似度，t 是阈值，当 C_s 与 C_t 的语义相似度大于 t 时，则将$<C_s, C_t>$作为等价概念映射对。

定义 2：本文认为《同义词词林》(扩展版)中收录的全部词汇以及它们之间的语义关系可构成一个语义知识库(Semantic Knowledge Base，SKB)，记做：SKB_{TYCCL}。显然集合 SKB_{TYCCL} 由原子概念组成，即有 $SKB_{TYCCL}=\{AC_1, AC_2,\cdots, AC_N\}$。$N$ 为知识库中所收录的词元总数。

定义 3：组合概念 CC_i 由一系列原子概念的有序

排列构成。对于 $\forall AC_i \in SKB_{TYCCL}$，引入二维下标 i 和 j，则有有序序列 $CC_i=[AC_{i1}, AC_{i2}, \cdots, AC_{ij}]$，其中 $j \geqslant 1$ 且 $CC_i \notin SKB_{TYCCL}$，j 为原子概念 AC_i 在有序序列 CC_i 中的排列位置。特别地，对于所有的原子概念 AC_i，可以有 $AC_i=[AC_i]$。

定义 4：对于本体 O_s 和 O_t 中的概念 C_s 和 C_t，有 $C_s=CC_s=[AC_{s1}, AC_{s2}, \cdots, AC_{sm}]$，$C_t=CC_t=[AC_{t1}, AC_{t2}, \cdots, AC_{tn}]$。$m$ 和 n 分别为概念 C_s 和 C_t 所对应的有序序列 CC_s 和 CC_t 的长度，则有 $m, n \geqslant 1$。

4　基于同义词词林和序列比对的中文关联数据模型

该模型主要由以下功能模块组成：本体预处理、组合概念分词处理、改进的同义词词林相似度计算、构建打分矩阵以及组合概念相似度计算(包含:改进的同义词词林相似度计算和序列比对处理)。系统总体框架如图 1 所示。基于上述形式化定义，将对中文本体概念映射过程中的各种情况进行分类讨论。

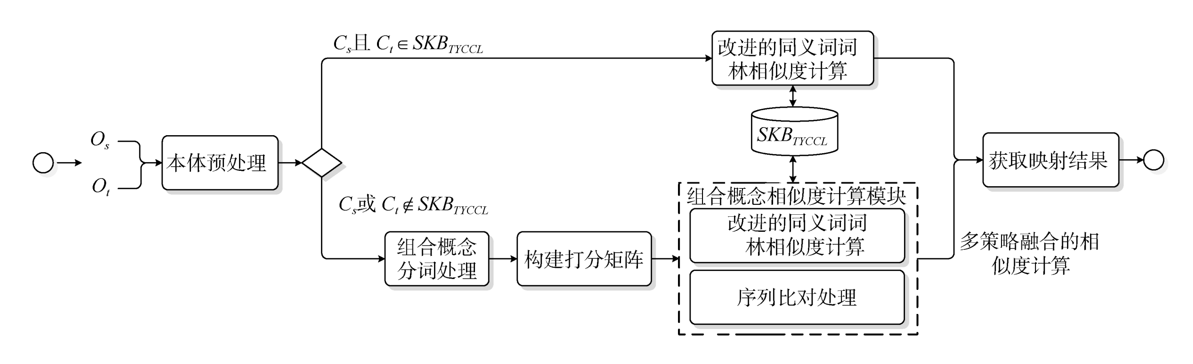


图 1　基于同义词词林和序列比对的中文本体映射模型

对于待映射的源本体 O_s 和目标本体 O_t 中的任意两个概念 C_s 和 C_t，在进行概念的语义相似度计算时，会出现如下三种情况：

(1) C_s 和 C_t 均为原子概念，即：$C_s \in SKB_{TYCCL}$ 且 $C_t \in SKB_{TYCCL}$；

(2) C_s 和 C_t 的其中之一为原子概念，而另一个为组合概念，即：$C_s \notin SKB_{TYCCL}$ 或 $C_t \notin SKB_{TYCCL}$；

(3) C_s 和 C_t 均为组合概念，即：$C_s \notin SKB_{TYCCL}$ 且 $C_t \notin SKB_{TYCCL}$。

对于情况(1)，本文直接采用“改进的同义词词林相似度计算 ”模块实现两个原子概念的语义相似度计算。

对于情况(2)和情况(3)的组合概念相似度计算的处理对策，本文将采用基于“序列比对处理 ”与“改进的同义词词林相似度计算 ”的多策略融合方式实现，即：“组合概念相似度计算 ”模块的输入为两个待映射的词串序列 CC_s 和 CC_t，以及其所对应的打分矩阵，该打分矩阵则由“组合概念分词处理 ”模块和“构建打分矩阵 ”模块协作生成。

4.1　基于改进的同义词词林相似度计算

同义词词林[27]是一个中文同义词典，它将每个词汇进行编码并以层次关系组织在一个倒挂的树形结构中，树中的每个节点代表一个概念，而中文的概念共指关系识别，实际上可以抽象为中文同义词的识别和语义相似度的计算问题，因此同义词词林是最佳的选择。本文采用哈尔滨工业大学同义词词林(扩展版)作为中文本体映射关系抽取的常识知识库。

在实验过程中，发现田久乐等[18]提出的传统算法过分强调概念之间的语义相关性，即：同义词词林中的层级之间的词汇父子类关系对于本体概念的等价关系获取会造成较大干扰。而本体映射任务却是要发现概念之间的等价关系而非父子类关系，因此本文通过引入语义调节因子和概念相似度权重系数来对传统算法进行改进，使之适用于LOD环境中的中文本体映射任务。

同义词词林将词元组织为分层结构，自顶向下共有 5 层。每个层次都有相应的编码标识，5 层的编码从左至右依次排列起来，形成词元的词林编码。词语与词语之间隐含的语义相关度也随着层次的增加而提高。

以词元“物质”为例(词林编码为：Ba01A02=)进行编码格式解释，如表 1 所示。

表 1　词林编码示例

编码位	1	2	3	4	5	6	7	8
子编码	B	a	0	1	A	0	2	"=(=或#或@)"
含义	大类	中类	小类		词群	原子词群		同义\不等\孤立
层次	第 1 层	第 2 层	第 3 层		第 4 层	第 5 层		

根据同义词词林的结构特点，首先对待映射概念的词林编码进行解析，抽取出第1至第5层子编码，然后从第 1 层子编码开始比较。若子编码不同，则根据所出现的层次来赋予该映射对相应的相似度权重。子编码不同的情况出现在越深的层次，则相似度权重越高；出现子编码不同的编码位越小(层次越浅)，其语义相关性就越差(相似度权重越低)。即：改进的方法可以同时兼顾词林中的层级因素对相似度计算结果的影响。同时，每层的分支节点数的多少也对相似度有影响。

本文给出基于同义词词林的相似度计算方法如公式(1)所示。

$$SIM_T(C_s,C_t)=\lambda\times\frac{L_i}{|L|}\times\cos(N_t\times\frac{\Pi}{180})\times\left(\frac{N_t-D+1}{N_t}\right)\tag{1}$$

由于本体映射任务更关注概念之间的语义相似性，因此需要通过引入语义调节因子 λ 来调节不同层级概念间语义相关性和语义相似性的关系以及控制处于不同层次分支的词元之间可能相似的程度，显然 $\lambda\in(0,1)$。λ 值越大，表示不同层次之间的词元相似或等价的可能性越大，且不同层次的语义相关性对于最终概念相似度的影响越大，反之则越小。特别地，在面对中文本体映射任务时，由于更突出概念间的语义相似度，因此 λ 取值不宜过高。

本文引入 $L=\{1,2,3,4,5\}$，对于 $\forall L_i\in L$，L_i 为子编码不同所出现在的层次数，$|L|$表示集合 L 中的元素个数，在本系统中恒等于 5。本文提出的概念相似度权重系数为 $\lambda\times(L_i/|L|)$。N_t 为词元 C_s 和 C_t 在第 i 层分支上的节点总数，D 为词元 C_s 和 C_t 的编码距离。特别地，当待映射概念对的 5 层编码均相等且词林编码最后一位为"="时，规定相似度函数 SIM_T 的返回值为 1.0。显然，函数 SIM_T 的值域为(0,1]。

4.2　基于序列比对的组合概念相似度计算

对于中文组合概念的相似度计算，许多学者提出了解决方案。例如：李佳等[14]设计并实现了基于知网(HowNet)的元素层概念相似度计算方法并实现了中文本体映射系统。该方法在处理未登录词的相似度计算问题时，将两个组合概念所对应的原子概念序列进行遍历，找出其中相似度最大的原子概念映射对，通过得到的相对极大的映射对求出两个组合概念的相似度值，如公式(2)所示。

$$Sim(A,B)=\frac{\sum_{i=1}^{\max(m,n)}\max_i(B_{xy})}{\max(m,n)}\tag{2}$$

其中，B_{xy} 表示分别以两个词汇拆分后得到的已知词为行列组成的相似度矩阵中的元素，$\max_i(B_{xy})$表示矩阵中数值排列为第 i 位的相似度。$\max(m,n)$表示取行号或列号的较大者。

但是，由于中文概念普遍存在"语序敏感"的特点，因此上述前人的处理方式难免带来语义相似度计算的误差。例如，不同本体中出现的两个待映射组合概念："历史理论"和"思想史"，经过分词处理后得到两个由原子概念构成的有序排列：[历史，理论]和[思想，史]。如果采用前人处理未登录词的普遍方法，则会得到如图 2 所示的原子概念映射结果。基于《同义词词林》(扩展版)并由公式(1)计算每对原子概念映射时的语义相似度，最后采用公式(2)进行综合计算得到的概念元素级相似度的值为 1.0，显然这是完全不合理的组合概念映射对和相似度结果。原因是该方法忽视了中文自然语言中普遍存在的"语序敏感"现象和"一词多义"的语义特征。

思想　　史
历史　　理论

图 2　错误的匹配结果

因此，本文提出一种改进的概念语义相似度计算方法。具体地，在计算概念之间的元素级相似度时，引入基于生物信息学的全局双序列比对算法进行语义相似度计算。

(1) 序列比对(Sequence Alignment)算法概述

在生物信息学中，双序列比对是指将两条 DNA、RNA 或蛋白质序列排列在一起，并标明其相似处。序列中可以插入空位符，对应的相同或相似的符号排在同一列上。通过比较两个序列间的相似片断和保守性位点，寻找其可能存在的分子进化关系[28]。

总体来说，比对模型可以分为两类：一类是全局

比对(Global Alignment)，主要考察两个序列之间的整体相似性，对序列进行全程扫描和比较。另一类是局部比对(Local Alignment)，重点关注序列中的某些特殊片断，比较序列中片断之间的相似性。二者均可通过动态规划(Dynamic Programming, DP)思想求解。

(2) 构造动态规划打分矩阵

所谓序列是指由一系列字母标识，根据一定的排列规则所组成的字符串。

①组合概念分词处理

本系统将组合概念视为词串序列，序列中的各个元素即为原子概念。将组合概念进行分词处理，得到其对应的词串序列，之后采用中国科学院计算技术研究所研发的ICTCLAS50[29]作为分词处理工具。字母表规定为《同义词词林》(扩展版)语义知识库：SKB_{TYCCL}。

②构建打分矩阵

首先将待比对的两个词串序列以打分矩阵 **M**(Scoring Matrix)的形式表示，两个序列分别作为动态规划矩阵的两维。对于待映射本体 O_s 和 O_t 中的概念 C_s 和 C_t，打分矩阵 **M** 的第 i 行对应词串序列 CC_s 中的原子概念 AC_{si}，第 j 列对应词串序列 CC_t 中的原子概念 AC_{tj}，其中 $I \leqslant m$，$j \leqslant n$。动态规划矩阵 **M** 中第 i 行第 j 列元素称为 M_{ij}。

根据动态规划思想，将两个词串序列以行和列来表示。假设序列 CC_s 的长度为 m，序列 CC_t 的长度为 n，则可形成一个以序列 CC_s 为行、序列 CC_t 为列的$(m+1)\times(n+1)$二维矩阵。例如：组合概念“第二次工业革命”和“第二次世界大战战犯”经过分词处理后，可以得到两个待比对词串序列：CC_s = [第二，次，工业革命]，CC_t = [第二，次，世界大战，战犯]。

(3) 最优化的递归求解算法

将本体映射的概念相似度计算抽象为两个词串序列的比对过程：通过空位罚分函数，决策在词串序列中的相应位置插入空位符“-”，使得两个序列长度相同，进而构建出待比对序列的原子概念之间或原子概念与空位符的对应关系。序列比对算法的本质就是通过评分策略，找出两个组合概念序列的最佳全局配对。

Needleman-Wunsch 算法于 1970 年由 Needleman 和 Wunsch 提出，是一种典型的用来比对序列之间全局相似性的动态规划算法，适用于比较全局宏观上相似程度较高的两个序列[30]。本文主要基于该算法和动态规划思想，对矩阵 M 中的最优比对路径进行递归求解。

算法 1: *ConceptSimilarity*(CC_s, CC_t)

输入：组合概念 CC_s 和 CC_t 所对应的打分矩阵 $M_{(i)(j)}$

输出：包含最优比对路径的矩阵 $M'_{(i)(j)}$

① $p \leftarrow -0.05$

//定义常量 p 为算法的惩罚因子，且等于–0.05

② for each $i \leftarrow 1, 2, \cdots, m+1; j \leftarrow 1, 2, \cdots, n+1$

//动态规划矩阵初始化

③ $M_{(i)(n+1)} \leftarrow p\times(m-i+1)$

④ $M_{(i+1)(j)} \leftarrow p\times(n-j+1)$

⑤ end for

⑥ for each $i \leftarrow m, m\text{-}1, \cdots, 1$

⑦ for each $j \leftarrow n, n\text{-}1, \cdots, 1$

⑧ $M_{(i)(j)} \leftarrow \max(M_{(i+1)(j+1)} + SIM_T(AC_{si}, AC_{tj}), M_{(i)(j+1)} + \text{p}, M_{(i+1)(j)} + \text{p})$ //递归计算矩阵中每个元素的代价值

⑨ end for

⑩ end for

⑪ 回溯得到包含序列比对最优路径的矩阵 $M'_{(i)(j)}$

⑫ return $M'_{(i)(j)}$

首先，给出序列比对算法的惩罚因子 p=–0.05，并分别对矩阵的第 n+1 列与第 m+1 行进行初始化。初始化规则分别为：$M_{(i)(n+1)} = \text{p}\times(m-i+1)$和 $M_{(m+1)(j)} = \text{p}\times(n-j+1)$。

其次，基于同义词词林相似度计算函数 SIM_T，对打分矩阵中其余的 $m\times n$ 个元素进行递归求解。本文给出记分函数 f 的定义，如公式(3)所示。

$$f(AC_{si}, AC_{tj}) = \begin{cases} SIM_T(AC_{si}, AC_{tj}) & if\ AC_{si} \neq "-" \text{且} AC_{tj} \neq "-" \\ f(AC_{si}, -) = \text{p} = -0.05 & if\ AC_{tj} = "-" \\ f(-, AC_{tj}) = \text{p} = -0.05 & if\ AC_{si} = "-" \end{cases} \quad (3)$$

考虑到中文组合概念普遍存在“词序敏感”的特点，将递归的起点选定为两个组合概念的结尾处，即：矩阵中的 M_{mn} 元素。对 SIM_T 的描述见公式(1)。递归规则(即：空位罚分函数)如公式(4)所示。

$$M_{ij} = \max\begin{cases} M_{(i+1)(j+1)} + f(AC_{si}, AC_{tj}) \\ M_{(i)(j+1)} + \text{p} \\ M_{(i+1)(j)} + \text{p} \end{cases} \quad (4)$$

最后，从矩阵中的 M_{mn} 元素开始，回溯至矩阵中的 M_{11} 元素结束，即可得到最优比对路径。在蕴含最优匹配路径的打分矩阵中，“加粗箭头”表示得到的最优路径。具体地，插入空位符“-”的策略为：“加粗斜箭头”表示将其尾部所对应的两个原子概念进行配对；“加粗水平箭头”表示对词串序列 CC_s 中，在其所在行对应的原子概念位置前插入一个空位符“-”；“加粗垂直箭头”表示对词串序列 CC_t 中，在其所在列对应的原子概念相应位置前插入一个空位符“-”。这里需要说明的是，如果得到的最优比对路径不止一条，则任选其一。

具体的基于全局序列比对思想的概念元素级相似度计算算法，见算法 1。

在插入空位符“-”后，两个待映射组合概念词条序

列的长度相等，称为 CC'_{si}和 CC'_{ti}，定义两组序列的长度为 L。最终根据比对结果，基于记分函数 f，得到组合概念之间的相似度计算方法如公式(5)所示。

$$SIM_{NW}(CC'_{si}, CC'_{ti}) = \sum_{i=1}^{|L|} \frac{f(AC_{si}, AC_{ti})}{|L|} \qquad (5)$$

5 实验数据与结果分析

5.1 数据来源

本文采用中文网络开放百科知识库作为实验数据源。除 DBpedia(中文版)知识库以外，本系统基于文献[5,31]，使用爬虫工具包 HTMLParser 分别对百度百科和互动百科的开放分类页面和词条页面所包含的 Infobox 结构化信息进行爬取和解析，并将其以中文三元组(Triple)的形式组织起来，形成待映射的大规模中文开放域知识库。如表 2 所示，本体概念体系主要由百科开放分类体系构成。

表 2 中文网络百科知识库信息

项目		百度百科	互动百科	DBpedia 3.8 (中文版)
本体概念	子分类	13	13	23
	中文三元组数量	1 323	29 263	106 000
Infobox 知识	Infobox 数量	214 732	257 215	204 822
	Infobox 中的谓词数	21 152	1061	18 206
	中文三元组数量	1 698 149	2 161 616	4 077 898
词条实例	Infobox 出现频率	2.30%	10.10%	19.74%
	中文三元组数量	9 346 184	2 545 447	1 037 557

5.2 评测指标

本文采用对中文概念等价关系识别的准确率(Precision)、召回率(Recall)和综合评价指标(F-measure)作为最终的评价标准。其中:

$$\text{Precision(P)} = \frac{\text{输出的正确映射对数}}{\text{输出的映射对总数}} \times 100\%$$

$$\text{Recall(R)} = \frac{\text{输出的正确映射对数}}{\text{标准结果中的映射对总数}} \times 100\%$$

$$\text{F-measure(F1)} = \frac{2 \times P \times R}{(P+R)} \times 100\%$$

笔者邀请首都经济贸易大学信息学院的 4 位本科四年级学生，采用人工识别和手工标注的方式对 DBpedia、百度百科和互动百科顶层分类树中客观存在的中文概念等价关系进行完整的获取，并以标注的结果作为本体映射实验的参考正确映射对，如表 3 至表 5 所示。

表 3 Baidu-Hudong 映射任务本体参考映射数统计

映射任务	顶层分类	Baidu 概念数量	Hudong 概念数量	参考映射对数
Baidu-Hudong	人物	120	1 497	57
	科学	157	2 323	62
	社会	102	3 937	60
	历史	118	2 093	54
	艺术	84	1 506	55
	自然	104	5 688	71
	体育	165	258	59
	地理	133	3 632	97

表 4 Hudong-DBpedia 映射任务本体参考映射数统计

映射任务	顶层分类	Hudong 概念数量	DBpedia 概念数量	参考映射对数
Hudong-DBpedia	人物	1 497	4 737	380
	科学	2 323	156	33
	社会	3 937	10 676	303
	历史	2 093	5 648	524
	艺术	1 506	1 908	193

表 5 Baidu-DBpedia 映射任务本体参考映射数统计

映射任务	顶层分类	Baidu 概念数量	DBpedia 概念数量	参考映射对数
Baidu-DBpedia	人物	120	4 737	26
	社会	102	10 676	28
	历史	118	5 648	11
	艺术	84	1 908	25
	地理	133	37 936	95

5.3 序列比对结果分析

在对基于序列比对的组合概念相似度计算方法进行阐述后，对之前提到的两组相似度计算算例进行重新审视。

算例 1: CC_s=[思想，史]，CC_t=[历史，理论]。由公式(2)得到的组合概念相似度值为 $Sim(CC_s, CC_t)$=(1.0+1.0)/2=1.0，而采用基于序列比对算法得到的组合概念序列对齐效果如图 3 所示，其对应的打分矩阵如图 4 所示。最终得到的组合概念相似度值应为 $SIM_{NW}(CC'_{si}, CC'_{ti})$=(–0.05+1.0–0.05)/3=0.3。该组示例映射对来自 Hudong-DBpedia 映射任务中的“历史”子任务。

思想 史 -
- 历史 理论

图 3 正确比对结果

	历史	理论	
思想	0.9	0.95	-0.1
史	0.95	0.17	-0.05
	-0.1	-0.05	0

图 4　算例 1 的打分矩阵

算例 2: CC_s=[第二, 次, 工业革命], CC_t=[第二, 次, 世界大战, 战犯]。若采用公式(2)计算组合概念相似度, 则会得到错误的相似度值: $Sim(CC_s, CC_t)$=1.0, 这是因为原子概念"次"存在"一词多义"现象。具体地, 词元"次"在《同义词词林》(扩展版)中有多个编码项, 其中 Dn04B03="编码项给出了两个原子词元"第二"和"次"为等价词元的判定。因此, 根据公式(2)会得到 4 组原子概念映射结果为1.0的情况, 分别是: <第二, 次>=1.0, <第二, 第二>=1.0, <次, 第二>=1.0, 以及<次, 次>=1.0。代入公式(2)有: $Sim(CC_s, CC_t)$=(1.0+1.0+1.0+1.0)/4=1.0。而根据基于序列比对的算法计算最终得到的组合概念相似度值应为 $SIM_{NW}(CC'_{si}, CC'_{ti})$=(1.0+1.0+0.18-0.05)/4=0.5325。通过算法 1 得到的包含最优匹配路径的打分矩阵 $M'_{(i)(j)}$如图 5 所示, 其所对应的最优化序列匹配结果如图 6 所示。该组示例映射对来自 Baidu-DBpedia 映射任务中的"历史"子任务。

	第二	次	世界大战	战犯	
第二	2.13	1.43	0.38	0.08	-0.15
次	1.08	1.13	0.43	0.13	-0.1
工业革命	0.03	0.08	0.13	0.18	-0.05
	-0.2	-0.15	-0.1	-0.05	0

图 5　算例 2 的打分矩阵

第二	次	工业革命	—
第二	次	世界大战	战犯

图 6　算例 2 的序列匹配结果

由此可以看出, 算例 1 和算例 2 中的两个组合概念之间并无等价关系。而传统方法却分别给出相似度均为 1.0 表示极高相似度的错误结论。相反地, 由算法 1 所得到的相似度值则更合理。考虑到中文概念普遍存在的"词序敏感"和"一词多义"现象时, 采用基于 Needleman-Wunsch 算法的全局比对算法, 可以有效规避以文献[14]为代表的传统方法可能带来的错误映射。同时, 在面对组合概念(即: 未登录词)映射时, 如果其所对应的词串序列中的原子概念的语义顺序基本相同, 算法 1 的效果则应与传统方法基本一致。综上所述, 基于全局序列比对的概念元素级相似度算法在面对大规模中文本体映射任务时, 比传统方法更具优势和合理性。

5.4　大规模中文本体映射结果分析

在上述算法思想的理论指导下, 本文以中文三大网络百科知识库为数据源, 面向大规模关联数据构建的实际应用场景, 对所提出的原型系统进行性能评测。完成三大映射任务后得到评测结果, 如表 6 至表 8 所示, 分别给出了采用 4 种不同的典型相似度计算算法所得到的准确率、召回率以及 F1 值。第一种算法为跨语言通用的编辑距离相似度算法[32], 第二种算法为传统的基于同义词词林的中文词语相似度计算算法[18], 第三种算法为李佳等提出的基于 HowNet 的中文词语相似度算法 ELOMC[14], 第 4 种算法为本文提出的中文概念综合相似度计算算法。

为了保证公平性, 本文将判定概念等价关系的相似度阈值统一设定为 $t = 0.9$。

表 6 为 Baidu-Hudong 本体映射任务的概念相似度计算结果, 可以看出, 本文系统的准确率均值分别高出传统的同义词词林算法和 ELOMC 算法 41%和 39%左右, 而召回率则高出编辑距离算法和传统的同义词词林算法平均约 13%和 2%左右, 并与 ELOMC 算法基本持平。在综合评价指标 F1 值上, 本系统分别高出编辑距离算法、传统的同义词词林算法和 ELOMC 平均约 8%, 23%和 20%。

表 7 为 Hudong-DBpedia 本体映射任务的概念相似度计算结果, 在准确率方面, 本系统分别高出编辑距离算法、传统的同义词词林算法和 ELOMC 算法平均约 1%、10%和 11%左右。召回率则高出编辑距离算法和传统的同义词词林算法平均约 6%和 1%左右, 并与 ELOMC 算法基本持平。本系统的综合评价指标 F1 值则分别高出编辑距离算法、传统的同义词词林算法和 ELOMC 算法平均约 3%, 6%和 6%。

表 8 为 Baidu-DBpedia 本体映射的概念相似度计算结果, 在处理该组任务时, 本系统的准确率均值分别高出传统的同义词词林算法和 ELOMC 算法 39%和 43%左右。召回率则高出编辑距离算法、传统的同义

表 6 Baidu-Hudong 映射任务评测结果

映射任务	顶层分类	编辑距离算法[32]			传统的基于同义词词林相似度算法[18]			ELOMC 算法[14]			本文系统		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baidu-Hudong	人物	1.000	0.526	0.690	0.657	0.772	0.710	0.943	0.877	0.909	0.980	0.877	0.926
	科学	1.000	0.742	0.852	0.422	0.790	0.551	0.387	0.774	0.516	0.982	0.774	0.866
	社会	1.000	0.567	0.723	0.640	0.800	0.711	0.622	0.850	0.718	0.915	0.717	0.804
	历史	1.000	0.611	0.759	0.607	0.630	0.618	0.607	0.685	0.643	1.000	0.685	0.813
	艺术	1.000	0.727	0.842	0.769	0.909	0.833	0.729	0.927	0.816	1.000	0.927	0.962
	自然	1.000	0.704	0.826	0.369	0.775	0.500	0.364	0.775	0.495	0.965	0.775	0.859
	体育	1.000	0.763	0.865	0.554	0.780	0.648	0.516	0.814	0.632	0.980	0.814	0.889
	地理	1.000	0.691	0.817	0.470	0.804	0.593	0.491	0.845	0.621	0.988	0.835	0.905
	平均值	**1.000**	**0.666**	**0.797**	**0.561**	**0.782**	**0.645**	**0.582**	**0.818**	**0.669**	**0.976**	**0.800**	**0.878**

表 7 Hudong-DBpedia 映射任务评测结果

映射任务	顶层分类	编辑距离算法[32]			传统的基于同义词词林相似度算法[18]			ELOMC 算法[14]			本文系统		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Hudong-DBpedia	人物	0.973	0.861	0.913	0.955	0.889	0.921	0.949	0.924	0.936	0.956	0.924	0.940
	科学	0.939	0.939	0.939	0.838	0.939	0.886	0.886	0.939	0.912	1.000	0.939	0.969
	社会	0.964	0.894	0.928	0.793	0.983	0.878	0.794	0.993	0.883	0.990	0.993	0.992
	历史	0.987	0.987	0.987	0.992	0.989	0.990	0.979	0.992	0.986	0.994	0.992	0.993
	艺术	0.984	0.938	0.960	0.823	0.990	0.899	0.740	0.995	0.849	0.989	1.000	0.994
	平均值	**0.969**	**0.924**	**0.946**	**0.880**	**0.958**	**0.915**	**0.870**	**0.969**	**0.913**	**0.986**	**0.970**	**0.978**

表 8 Baidu-DBpedia 映射任务评测结果

映射任务	顶层分类	编辑距离算法[32]			传统的基于同义词词林相似度算法[18]			ELOMC 算法[14]			本文系统		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baidu-DBpedia	人物	1.000	0.692	0.818	0.452	0.731	0.559	0.733	0.846	0.786	0.963	1.000	0.981
	社会	1.000	0.679	0.809	0.639	0.821	0.719	0.500	0.821	0.622	0.955	0.750	0.840
	历史	1.000	0.727	0.842	0.692	0.818	0.750	0.188	0.818	0.305	0.900	0.818	0.857
	艺术	1.000	0.760	0.864	0.639	0.920	0.754	0.821	0.920	0.868	1.000	1.000	1.000
	地理	1.000	0.853	0.920	0.421	0.979	0.589	0.413	1.000	0.585	0.989	0.989	0.989
	平均值	**1.000**	**0.742**	**0.851**	**0.569**	**0.854**	**0.674**	**0.531**	**0.881**	**0.633**	**0.961**	**0.912**	**0.934**

词词林算法和 ELOMC 算法平均约 17%、6%和 3%左右。在综合评价指标 F1 值上，本系统分别高出编辑距离算法、传统的同义词词林算法和 ELOMC 算法平均约 8%, 26%和 30%。

而在 Baidu-Hudong 和 Baidu-DBpedia 映射任务中，本系统的准确率低于编辑距离算法，这是因为《同义词词林》(扩展版)中客观存在一些有争议的或是被不当归类为同义词对的情况。如果它们出现在结果集中，本文则视其为错误的映射结果，例如: <民族，中华民族>、<刑法，刑事>、<军队，军事>、<辛亥革命，革命>等。这种情况在该组映射的"社会"子映射任务中出现的次数较多，但是在其他映射任务中则较少出现有争议的同义词对。

从宏观上讲，本文模型在三大映射任务的总计 18 组子映射任务上获得的准确率、召回率和综合评价指标的平均值可以分别达到约 97.5%、87.8%和 92.1%。虽然本文模型在准确率上略低于编辑距离算法，但这是由于同义词词林中个别被不恰当归类的同义词对所

造成的；而编辑距离算法却只能单纯机械地比较概念之间的字面相似度，这种完全忽视概念之间的语义相似性的算法必然会导致其在所有映射任务中的召回率均明显低于其他系统。而本文方法由于引入语义词典——《同义词词林》(扩展版)，并对传统的基于同义词词林算法加以改进，因此在召回率上会明显高于编辑距离算法。这也就使得在最终综合评价指标 F1 值的比较上，本文方法在三组映射任务中均明显高于编辑距离算法。

综上所述，本文模型的综合评价指标为同类系统中最优；其准确率明显高于传统的同义词词林算法和 ELOMC 系统；而其召回率则高于编辑距离算法和传统的同义词词林相似度算法，并与 ELOMC 系统基本持平。

6 结 语

现阶段缺乏成熟的中文大规模本体映射系统，本文针对关联数据网络构建过程中的本体模式匹配问题，提出一种新的基于同义词词林和全局序列比对算法相融合的中文本体映射模型。该系统解决了大规模本体映射系统的可用性问题。它着眼于现有中文大规模本体的"语序敏感"和"一词多义"特征，进行组合概念的元素级映射。今后将根据不同中文本体的特征，考虑引入实例级以及概念定义相似度的映射参数，进一步提高中文映射系统的健壮性和准确性。

参考文献:

[1] Berners-Lee T, Hendler J, Lassila O. The Semantic Web[J]. Scientific American, 2001, 284(5): 28-37.

[2] Borst W N. Construction of Engineering Ontologies for Knowledge Sharing and Reuse [D]. Universiteit Twente, 1997.

[3] Bizer C, Heath T, Idehen K, et al. Linked Data on the Web[C]// Proceedings of the 17th International Conference on World Wide Web, Beijing, China. New York, USA: ACM, 2008: 1265-1266.

[4] Niu X, Sun X, Wang H, et al. Zhishi. me-Weaving Chinese Linking Open Data[C]//Proceedings of the 10th International Conference on the Semantic Web, Bonn, Germany. Heidelberg, Germany: Springer-Verlag Berlin, 2011: 205-220.

[5] Wang Z, Wang Z, Li J, et al. Knowledge Extraction from Chinese Wiki Encyclopedias [J]. Journal of Zhejiang University-Science C: Computer & Electronics, 2012, 13(4): 268-280.

[6] Jain P, Hitzler P, Sheth A P, et al. Ontology Alignment for Linked Open Data [C]//Proceedings of the 9th International Conference on the Semantic Web, Shanghai, China. Heidelberg, Germany: Springer-Verlag Berlin, 2010: 402-417.

[7] Melnik S, Garcia-Molina H, Rahm E. Similarity Flooding: A Versatile Graph Matching Algorithm and Its Application to Schema Matching[C]// Proceedings of the 18th IEEE International Conference on Data Engineering, San Jose, California, USA. Washington, USA: IEEE Computer Society, 2002: 117-128.

[8] Cohen W, Ravikumar P, Fienberg S. A Comparison of String Metrics for Matching Names and Records[C]// Proceedings of KDD Workshop on Data Cleaning and Object Consolidation. 2003, 3: 73-78.

[9] Giunchiglia F, Yatskevich M. Element Level Semantic Matching [C]//Proceedings of Meaning Coordination & Negotiation Workshop at ISWC. 2004.

[10] Stark M M, Riesenfeld R F. WordNet: An Electronic Lexical Database[C]//Proceedings of the 11th Eurographics Workshop on Rendering. MIT Press, 1998.

[11] Isaac A, Van Der Meij L, Schlobach S, et al. An Empirical Study of Instance-Based Ontology Matching[C]//Proceedings of the 6th International the Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference.Heidelberg, Germany: Springer-Verlag Berlin, 2007: 253-266.

[12] Nikolov A, Uren V, Motta E, et al. Integration of Semantically Annotated Data by the KnoFuss Architecture [C]// Proceedings of International Conference on Knowledge Engineering and Knowledge Management. Heidelberg, Germany: Springer-Verlag Berlin, 2008: 265-274.

[13] Zhong Q, Li H, Li J, et al. A Gauss Function Based Approach for Unbalanced Ontology Matching [C]// Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. ACM, 2009: 669-680.

[14] 李佳，祝铭，刘辰，等. 中文本体映射研究与实现[J]. 中文信息学报，2007, 21(4): 27-33. (Li Jia, Zhu Ming, Liu Chen, et al. Research and Implementation on Chinese Ontology Mapping [J]. Journal of Chinese Information Processing, 2007, 21(4): 27-33.)

[15] 董振东，董强，郝长伶. 知网的理论发现[J]. 中文信息学报，2007, 21(4): 3-9. (Dong Zhendong, Dong Qiang, Hao Changling. Theoretical Findings of HowNet[J]. Journal of

Chinese Information Processing, 2007, 21(4): 3-9.)

[16] 陆丙甫. 语序优势的认知解释[J]. 当代语言学, 2005, 7(1): 1-15. (Lu Bingfu. Word Order Dominance and Its Cognitive Explanation [J]. Contemporary Linguistics, 2005, 7(1): 1-15.)

[17] 哈尔滨工业大学社会计算与信息检索研究中心. 同义词词林(扩展版)[EB/OL]. [2014-09-05]. http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm. (HIT-SCIR. TongYicCi CiLin (Extended Version) [EB/OL]. [2014-09-05]. http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm.)

[18] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报: 信息科学版, 2010, 28(6): 602-608. (Tian Jiule, Zhao Wei. Words Similarity Algorithm Based on Tongyici Cilin in Semantic Web Adaptive Learning System [J]. Journal of Jilin University: Information Science Edition, 2010, 28(6): 602-608.)

[19] Volz J, Bizer C, Gaedke M, et al. Silk-A Link Discovery Framework for the Web of Data[C]// Proceedings of LDOW2009, Madrid, Spain. 2009.

[20] Volz J, Bizer C, Gaedke M, et al. Discovering and Maintaining Links on the Web of Data[C]//Proceedings of the International Semantic Web Conference. Springer Berlin Heidelberg, 2009: 650-665.

[21] Hassanzadeh O, Lim L, Kementsietsidis A, et al. A Declarative Framework for Semantic Link Discovery over Relational Data[C]//Proceedings of the 18th International Conference on World Wide Web. ACM, 2009: 1101-1102.

[22] Baidu Baike [EB/OL]. [2015-09-10]. http://baike.baidu.com/.

[23] Hudong [EB/OL]. [2015-09-17]. http://www.hudong.com/.

[24] DBpedia [EB/OL]. [2015-09-30]. http://wiki.dbpedia.org/.

[25] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia-A Crystallization Point for the Web of Data [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2009, 7(3): 154-165.

[26] Wang Z, Li J, Tang J. Boosting Cross-Lingual Knowledge Linking via Concept Annotation[C]// Proceedings of the International Joint Conference on Artificial Intelligent. 2013.

[27] 梅家驹, 竺一鸣, 高蕴琦, 等. 同义词词林[M]. 上海: 上海辞书出版社, 1983. (Mei Jiaju, Zhu Yiming, Gao Yunqi, et al. TongYiCiCiLin[M]. Shanghai: Shanghai Lexicographical Publishing House, 1983.)

[28] Setubal J C, Meidanis J. Introduction to Computational Molecular Biology [M]. PWS Pub.Co., 1997.

[29] 中国科学院计算技术研究所. ICTCLAS [EB/OL]. [2013-01-03]. http://ictclas.org/. (NLPIR Chinese Word Segmentation System [EB/OL]. [2013-01-03]. http://ictclas.nlpir.org/.)

[30] Needleman S B, Wunsch C D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of two Proteins [J]. Journal of Molecular Biology, 1970, 48(3): 443-453.

[31] Wang T, Song J, Di R, et al. A Thesaurus and Online Encyclopedia Merging Method for Large Scale Domain-Ontology Automatic Construction [C]//Proceedings of the International Conference on Knowledge Science, Engineering and Management. Heidelberg, Germany: Springer-Verlag Berlin, 2013: 132-146.

[32] Levenshtein V I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals [J]. Soviet Physics Doklady, 1966, 10: 707-710.

作者贡献声明:

王汀: 提出研究方法、思路并进行系统设计与实现、论文起草;
高迎: 论文修改;
刘经纬: 参与研究思路的设计, 数据分析。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 王汀. HIT-IRLab-同义词词林(扩展版)_full_2005.3.3.txt. 哈工大同义词词林(扩展版).

[2] 王汀. ICTCLAS50_Windows_32_JNI.rar. NLPIR 汉语分词系统.

[3] 王汀. Baidu-Ontology-Concepts.rar. 百度百科顶层分类树-13 大类本体概念集.

[4] 王汀. Hudong-Ontology-Concepts.rar. 互动百科顶层分类树-13 大类本体概念集.

[5] 王汀. DBpedia V3.8 zh-(子本体).rar. 中文维基百科顶层分类树-23 大类本体概念集.

收稿日期: 2016-08-18
收修改稿日期: 2016-10-22

Linking Chinese Open Data at Schema-Level

Wang Ting　Gao Ying　Liu Jingwei
(Information School, Capital University of Economics and Business, Beijing 100070, China)

Abstract: [**Objective**] This study proposes a novel Chinese Ontology Mapping model based on TongYiCiCiLin (TYCCL) and Sequence Alignment to evaluate concept similarity of the Linked Chinese Open Data at Schema-Level. [**Methods**] Firstly, we modified the TYCCL-based algorithm to compute the similarity among atomic Chinese concepts from the TYCCL. Secondly, we proposed a global sequence-alignment algorithm to evaluate the similarity among Chinese OOV. [**Results**] The proposed model was examined with the corpus from DBpedia (Chinese version), Baidu baike and Hudong knowledge base. The Precision, Recall and F1-value of this model were 97.5%, 87.8% and 92.1%, respectively. [**Limitations**] The proposed model only measured the similarity among Chinese Ontology concepts at the element level, which did not evaluate the impacts of Ontology attributes and instance on the concept equivalence relationship. [**Conclusions**] The proposed model is better than existing ones.

Keywords: Chinese Linked Open Data　TongYiCiCiLin　Sequence Alignment　Ontology Mapping　Similarity Computing

---

NISO 发布 ResourceSync(资源同步)框架规范的更新版本

美国国家信息标准组织(NISO)于近日宣布正式出版了 ResourceSync 框架规范的更新版本(ANSI / NISO Z39.99-2017)。由美国国家标准协会(ANSI)批准，该 1.1 版本改进了一个 Web 标准，该标准详细说明了服务器可以实现的各种功能，以允许第三方系统与不断发展的资源保持同步。这种同步在当前的环境下是非常重要的，现如今，不仅是内容的元数据，基于 Web 的内容也在不断变化。

ResourceSync 在 2014 年首次发布 ANSI / NISO Z39.99。该标准也称为 ResourceSync“核心”规范，提供了一系列易于服务器实现的功能，以使远程系统与不断发展的资源保持更紧密地同步。它还描述了服务器应如何声明其支持的设施，并提供大量的示例和用例来指导用户实施。最近的修订版修正了资源的最新修改日期与资源修改的通知日期的混淆等相关问题。

“Web 资源和 Web 资源集合不断发展，在许多情况下，希望利用这些资源的应用程序需要确信他们使用的数据是最新的。” ResourceSync 工作组联合主席 Herbert Van de Sompel 说: “我们对 ResourceSync 核心规范的修订加强了一个标准，可以满足学术交流、文化遗产和教育等领域中不同系统之间的资源发现和同步需求。ResourceSync 在设计上非常模块化，基于 HTTP 和 Sitemap 协议，以确保在许多应用程序中能够轻松实现，包括但不限于及时共享来自不同类型的存储库的数据。此外，相关的可选规范提供了对 ANSI / NISO ResourceSync 核心的扩展，包括支持同步信息存档和基于推送的变更通知等规范。”

有关使用 ANSI / NISO Z39.99-2017 标准的 ResourceSync 规范和视频教程，请访问 NISO 网站 http://www.niso.org/workrooms/resourcesync/。

(编译自: http://www.niso.org/news/pr/view?item_key=96962d7722cc13a1e20c40e2ca3c2ca8ca80359d)

(本刊讯)